Case Study 4:
Calibrating Snow Gauge

By:
Himanshu Makhija (A09845605) -- Math/Computer Science, B.S.
Hanna Goldman () -- Math/Computer Science B.S.
Shagun Gupta (A91068956) -- Bioinformatics B.S.
Amey Paranjape (A53218045) -- ECE M.S.

MATH 189/289C- Winter 2017

The final code can be found at: https://github.com/hjgoldma/Math189HW3

Table of Contents

Introduction

Background

The main source of water for Northern California comes from the Sierra Nevada mountains. In order to determine the depth profile of the snow density, the Forest Service of the United States Department of Agriculture operates a gamma transmission snow gauge in Soda Springs, CA. Note that the gauge does not directly measure the density of the snow, but rather it is calculated from a measurement of gamma transmissions.

An important feature of the snow gauge is that it does not disturb the snow whilst performing the measurement, which allows researchers to run the process as many times as needed. By being able to run the measurement process repeatedly, researchers are able to study the effects of rain on snow throughout the season. When rain falls on snow, the snow density determines the amount of rain absorbed into the snow, where the denser the snow the less it can absorb. Flooding occurs when the rain is not able to be absorbed due to the density of the snow.

The goal is to analyze the snowpack profile in order to determine its effects on the water supply and food management. However, due to instrument wear and the decay of radioactive source, the measurements are variable and result in modifying the function of converting the gain measurements into density. As a result, a calibration run is performed at the start of the winter season.

Goal

We are going to develop a procedure to transform the gain measurements into density measurement of the snow.

Description of Data

The dataset in this lab is based comes from the middle 10 of 30 measurements for each of 9 densities in g/cm^3 of polyethylene, which are used to simulate the snow densities but the densities of the polyethylene blocks are already known. Here, density is a numerical continuous variable and gain is also a numerical continuous variable.

But like most things in life, it is not as simple to find these measurements for gain. One thing that affects our knowledge base is a decline on operational networks in the northern regions (including Alaska, northern Canada, and Siberia). Plus, there are only a few stations in

the mountainous regions, which means we don't have a lot of sources of data to cross-compare. Another factor is biases in the gauge measurements across operational networks due to methods of data processing.

Snow Gauges

To get a sense of the data, it will also prove important to better understand what the snow gauge is. Due to its complexity and cost, it is not feasible to establish a broad range of snow gauges in a watershed area, and so the snow gauge is to be used primarily as a research tool rather than a method of monitoring a water supply for a specific area. The gauge in California is located at the center of a forest opening at roughly 2000 meters elevation and therefore is subject to all major high-altitude storms, that regularly deposit around 5-20 cm of wet snow. In this particular area, the average depth of the snow is 4m every winter.

The snow gauge itself consists of a Cesium-137 radioactive source and an energy detector mounted on separate vertical poles that are set a specific distance apart. The lift mechanism at the top of the poles raises and lowers the source and detector together, where the source emits gamma photos in all directions allowing the detector to count the photons eating through the 70-cm gap from source to detector. These signals are transmitted by a cable to the lab where the signal is stabilized and then converted to a measurement termed "gain", which should be directly proportional to the emission rate. The average range for snowpack density typically ranges between 0.1-0.6 g/cm^3.

Basically, the gamma ray on route to the detector crystal passes through a number of polyethylene molecules, which is determined by the density of the polyethylene. The molecule can either absorb the gamma photon, it may bounce it out of the path to the detector, or it can pass right through the molecule. Therefore, if each molecule acts independently, then the probability that a gamma photon successfully arrives at the detector is p^m, where p is the chance a single molecule will neither absorb nor deflect the ray, and m is the number of molecules in a straight line from the source to detector (determined by the density of the polyethylene block.

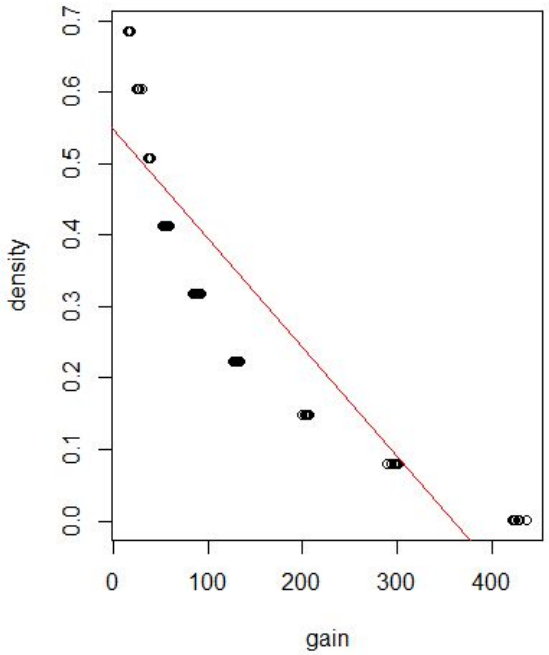
Hypothesis

Our hypothesis is that there exists a negative correlation between density and gain since greater is the density (more packed the ice is), more are the chances of water flooding or a lower gain.

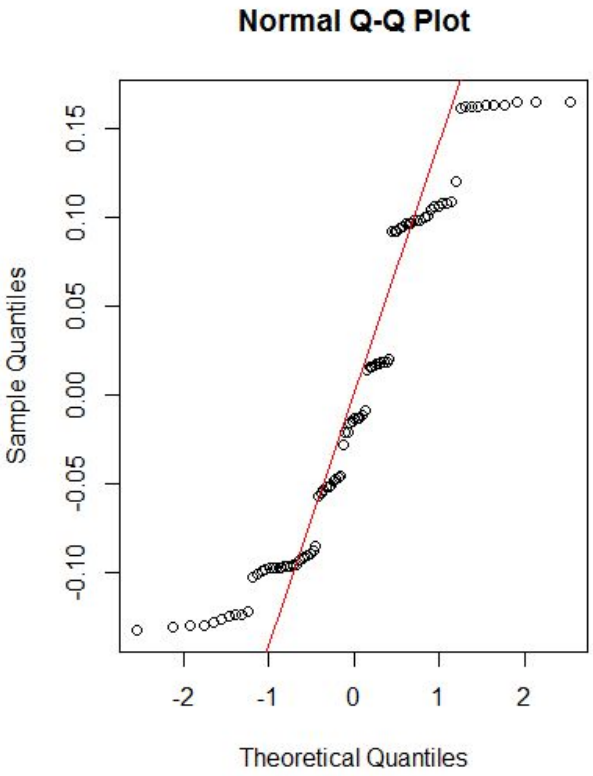
Body

Linear Fit

What we want from this study is to convert the gain measurements into density, and the first step would be to fit the data (the gain, or a transformation of the gain) through a least squares line with a scatter plot.



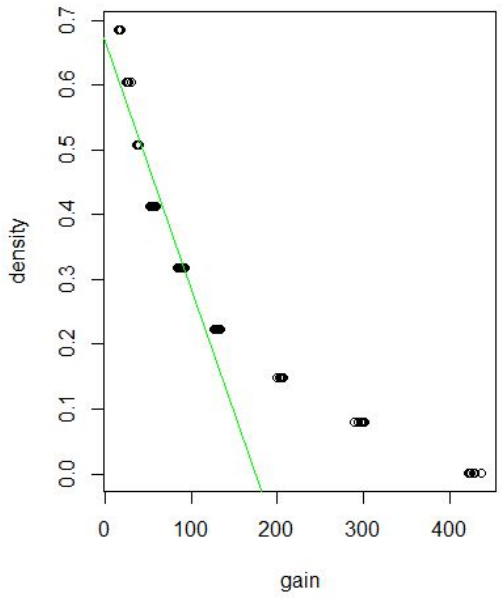
From the above, we can see the trend for gain against density does not seem to follow the linear line of fit. To confirm, we check the residuals plot for the above least squares line of fit.



By plotting the residuals from the linear fit, we can observe that the center points do follow the qqline closely near the middle, however note the ends and how far they stray from the line. As such, we move forward and try a quadratic line of fit to see if it fits our measurements for gain against density closer.

Quadratic

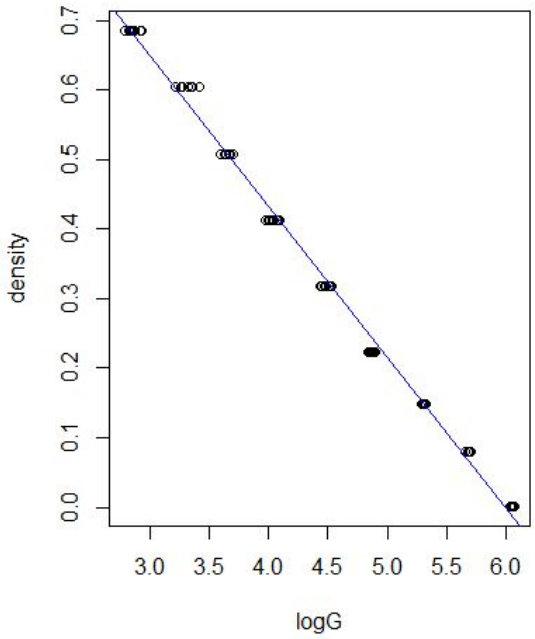
The next goal is to check whether or not a quadratic least squares fit provides us with a better approximation of the gain against density plot.



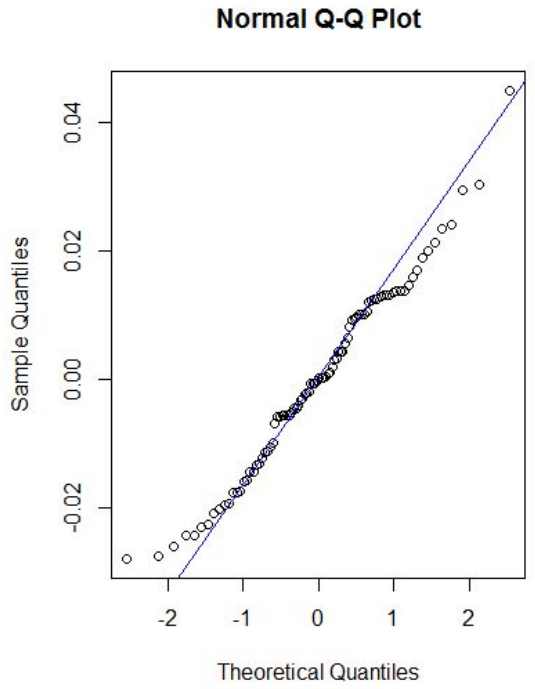
The line fails to closely approximate from our initial findings, and so we move onto another possibility.

Log Transform

Finally, the best results were obtained when we performed a log transform of the given gain measurements. This resulted in a shifting of the actual plot, over trying to find a better least squares approximation for the data as given.



As we can see, the least squares line of fit follows very closely to the transformed gain against density plot.



Now the residuals are closely following the qqline near the lower middle, but again stray near either ends. To close our findings for the best least squares line for the given gauge data, we use r^2, or the coefficient of determination.

Summary of Linear Fit:

```
Call:
lm(formula = density ~ gain, data = gauge)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13198 -0.09452 -0.01354  0.09682  0.16495

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5497239  0.0151243   36.35   <2e-16 ***
gain        -0.0015334  0.0000777  -19.73   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09769 on 88 degrees of freedom
Multiple R-squared:  0.8157,  Adjusted R-squared:  0.8136
F-statistic: 389.5 on 1 and 88 DF,  p-value: < 2.2e-16
```

Summary of Quadratic Fit:

```
Call:
lm(formula = density ~ gain + I(gain^2), data = gauge)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07086 -0.04343 -0.02213  0.03776  0.08513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.708e-01  1.110e-02   60.45   <2e-16 ***
gain        -3.844e-03  1.549e-04  -24.82   <2e-16 ***
I(gain^2)    5.499e-06  3.557e-07   15.46   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05076 on 87 degrees of freedom
Multiple R-squared:  0.9508,  Adjusted R-squared:  0.9497
```

F-statistic: 840.7 on 2 and 87 DF, p-value: < 2.2e-16

Summary of Log Transformed Gain Fit:
Call:
lm(formula = density ~ logG, data = gauge)

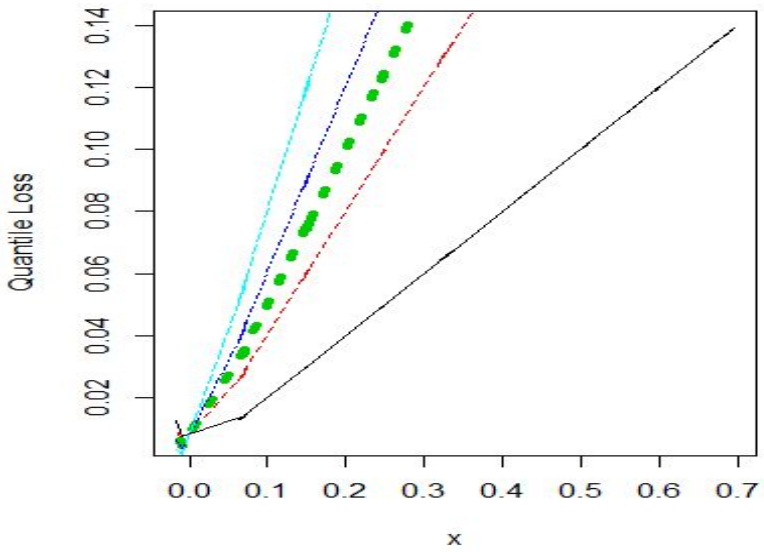
Residuals:
      Min   1Q   Median   3Q   Max
-0.028031 -0.011079 -0.000018 0.011595 0.044911

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.298013   0.006857   189.3   <2e-16 ***
logG   -0.216203   0.001494 -144.8   <2e-16 ***

Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01471 on 88 degrees of freedom
Multiple R-squared: 0.9958, Adjusted R-squared: 0.9958
F-statistic: 2.096e+04 on 1 and 88 DF, p-value: < 2.2e-16

Quantile Regression



```
        10%         30%         50%         70%         90%
-0.008468512  0.153587497  0.329386482  0.504061908  0.664969139
```

Confidence Interval for Slope

Recall that the confidence interval is calculated as a point estimate plus/minus the ME, where the degrees of freedom associated with the slope in a simple linear regression is (n-2). As such, with n = 90, and therefore the df = 88, we get a confidence interval of [-0.2186849, -0.2137177].

Cross-Validation

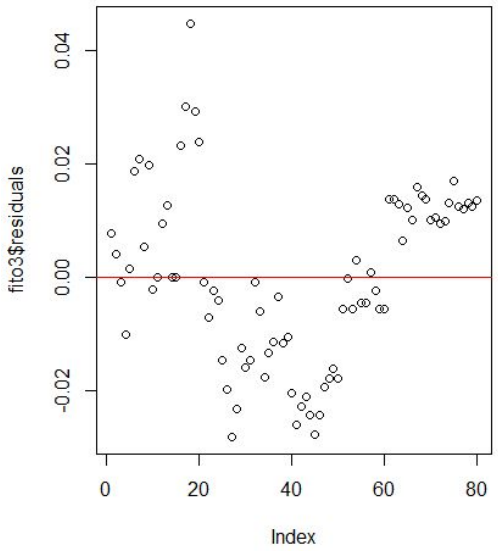
First, we removed all data points in the original gauge measurements that correspond to the 0.508 density. The summary of the quadratic least squares line of fit for our data changes as follows:

```
Call:
dataout1 <- data[data$density != 0.508,]
lm(formula = density ~ gain + I(gain^2), dataout1)

Coefficients:
(Intercept)         gain    I(gain^2)
  6.790e-01   -3.907e-03    5.603e-06
```

Similarly, the results for the gauge measurements without the 0.508 density with the log transform has its summary as follows:

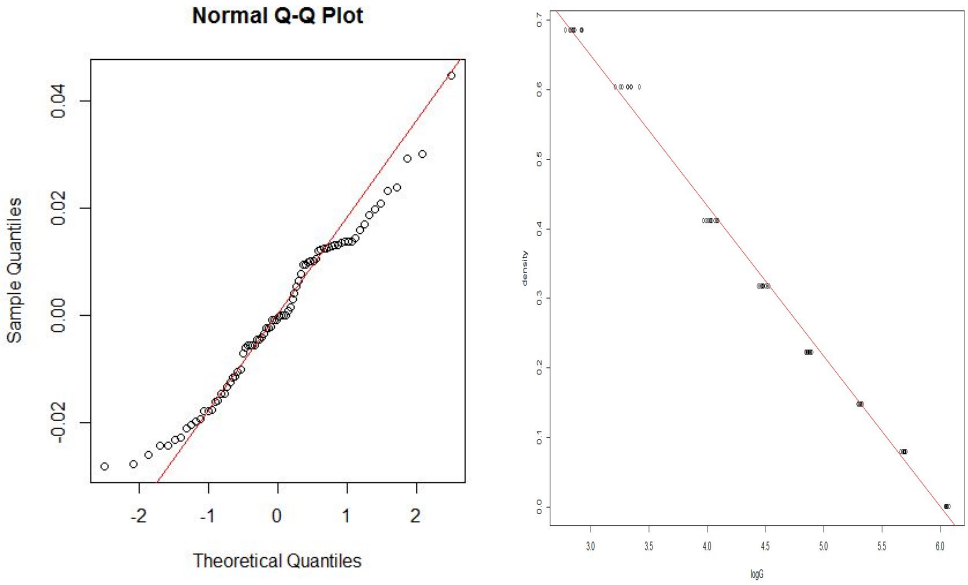
```
Call:
lm(formula = density ~ logG, data = dataout1)

Coefficients:
(Intercept)         logG
     1.2984      -0.2163
```

The residuals plotted for the log transform, since that is best fit we found, are as follows:



The residual plot gave us some variability since the points were found to be distributed over both sides of the red line at a given index point indicating constant variability.

To check how close the approximation is, we use a qqplot.

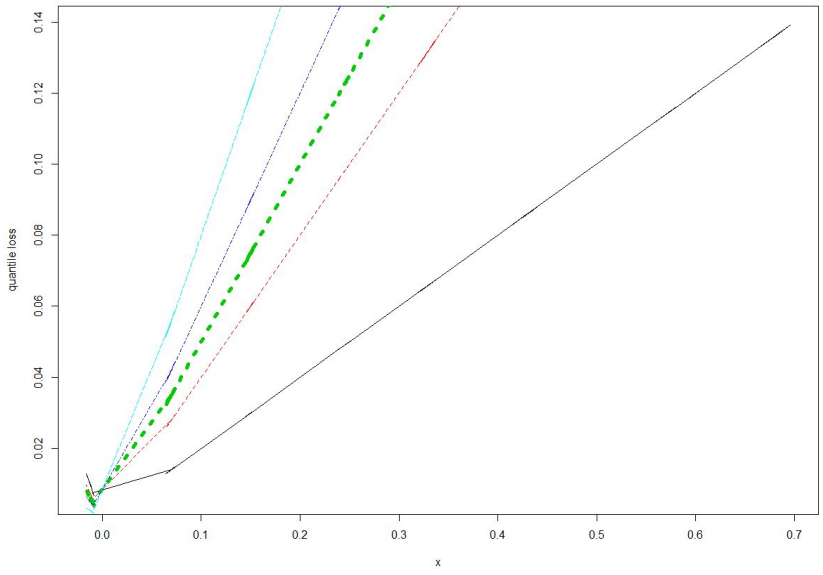


The Q-Q plot showed the points in this model lying along the expected normal line towards the middle thereby fulfilling the conditions for a least squared line to be a good fit for the log transformed model after removal of the outlier points (density = 0.508).

Call:

lm(formula = density ~ logG, data = dataout1)

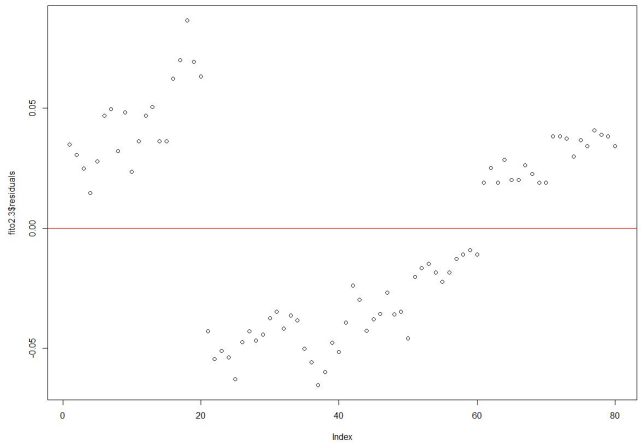
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.028143 -0.011781 -0.000524  0.012528  0.044757

Coefficients:
             Estimate   Std. Error   t value Pr(>|t|)
(Intercept)  1.298422   0.007679   169.1   <2e-16 ***
logG        -0.216278   0.001635  -132.2   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01547 on 78 degrees of freedom
Multiple R-squared:  0.9956,  Adjusted R-squared:  0.9955
F-statistic: 1.749e+04 on 1 and 78 DF,  p-value: < 2.2e-16
```



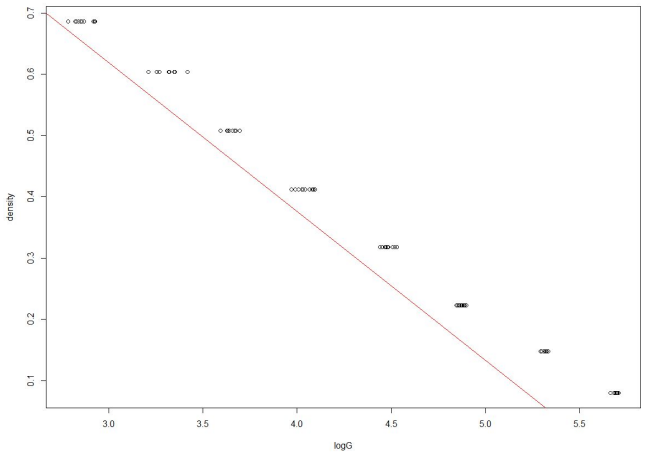
As expected, the quantile regression plot yielded a plot that was quite similar to the one we got before the removal of the outlier points (density = 0.508) since quantile regression is dependent on the median which shouldn't be really affected by the presence or absence of outliers. This can be contrasted the linear regression line which changed since it is dependent on the mean and hence affected by outliers.

We do this process once more, except this time only the densities corresponding to 0.001 were removed.

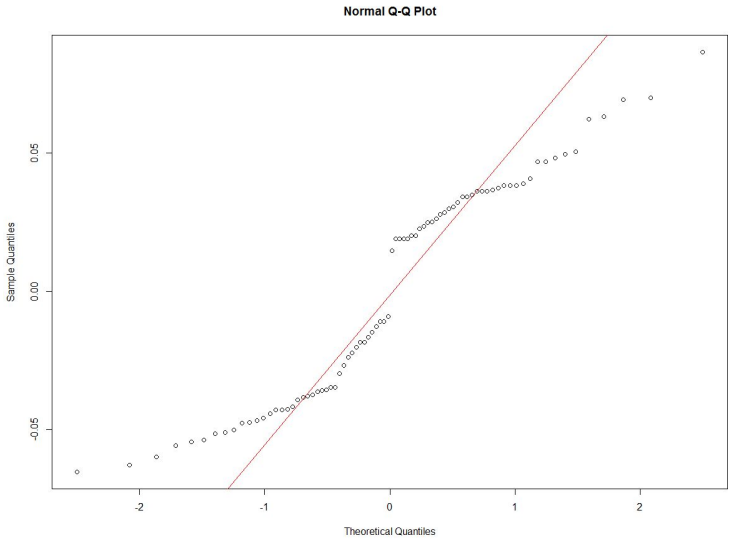
```
Call:
dataout2 <- data[data$density != 0.001,]
lm(formula = density ~ logG, data = dataout2)

Coefficients:
(Intercept)         logG
     1.3490      -0.2433
> summary(fito2.3)

Call:
lm(formula = density ~ logG, data = dataout2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065409 -0.038081  0.002694  0.035174  0.086598

Coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  1.349014   0.021370   63.13   <2e-16 ***
logG          -0.243323   0.004886  -49.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04054 on 78 degrees of freedom
Multiple R-squared:  0.9695,  Adjusted R-squared:  0.9691
F-statistic:  2480 on 1 and 78 DF,  p-value: < 2.2e-16
```



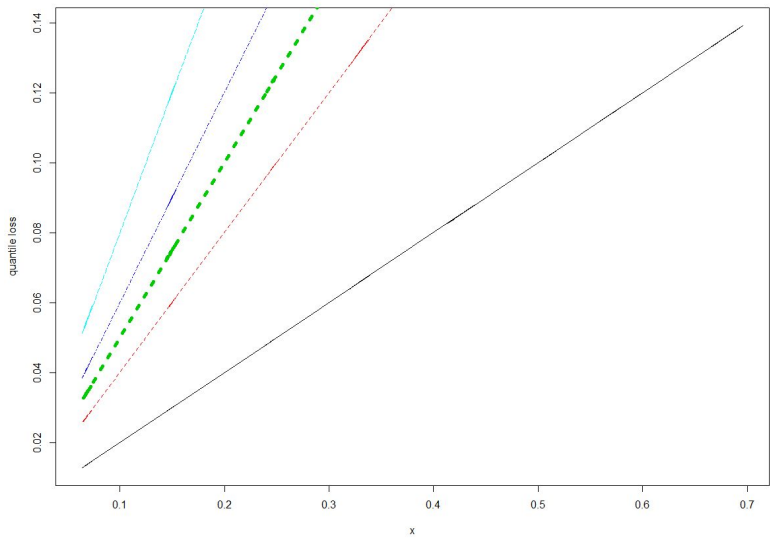
Our model did not show constant variability for the residual plot of the dataset after removal of all densities of 0.001. This indicates that the model fails the normal residuals test for linear regression model.



However the linear regression line for the scatter plot of our dataset gives a good linear plot as can be seen from the plot above.



This is further implied in the Q-Q plot since our log transformed data (gain) versus density does not lie along the red line and hence is not normally distributed.



As expected, the quantile regression plot is different than what we get for the entire dataset. This is because as was previously established that the log transformed model is not a good fit for a linear model.

Mathematical basis:

Confidence Interval for the log(Gain) (removing ten data point) after removal of all density points = 0.508 was for
n = 80 df = 80 − 2 = 78
> qt(.95, df=78)
[1] 1.664625

CI = -0.216278 ± 1.664625 × 0.001635 = (-0.2189,-0.2136)

The actual gain associated with this density was based on the linear fit of our model which was:
Linear regression line:
(Intercept) logG
1.2984 -0.2163
Density = 1.2984 + (-0.2163*(log(Gain)))
0.508 = 1.2984 + (-0.2163*(log(Gain)))
log(Gain) = 3.654184 => Gain = 4510.0775
This implies that our model does not work after cross validation since our expected log(Gain) value lies beyond that of our confidence interval after removal of all data points with density of 0.508

Confidence Interval (removing ten data point) after removal of all density points = 0.001 was for
n = 80 df = 80 − 2 = 78
> qt(.95, df=78)
[1] 1.664625

CI = -0.243323 ± 1.664625 × 0.004886 = (-0.2514,-0.2352)

The actual gain associated with this density was based on the linear fit of our model which was:
Linear regression line:
(Intercept) logG
  1.3490 -0.2433
Density = 1.3490 + (-0.2433*(log(Gain)))
0.001 = 1.3490 + (-0.2433*(log(Gain)))
log(Gain) = 5.5405 => Gain = 347,124.285

This implies that our model does not work after cross validation since our expected log(Gain) value lies beyond that of our confidence interval after removal of all data points with density of 0.001.

Conclusion

In conclusion, we followed the procedure to find the best least squares line of fit, in which linear, quadratic, and a log transform were attempted. The first two strayed far from our model, whereas the log transform showed a lot closer correlation. However, after removing the series of data points, the new results did not fit the confidence interval generated. This is to be expected because it is likely that there is no perfect model.

Theory

1. **Correlation**:

Before there is any fitting of bivariate data to any method, it is important to graph the data in a scatterplot to determine a rough estimate of which model to use. If the data seems to relate in some sort of linear fashion, linear fitting may seem to be a good idea. In order to support this notion, correlation may be calculated. Correlation is a measure of a systematic linear relationship between two sets of data. As long as there is a systematic relationship, the strength may vary between positive and negative 1. The correlation may be calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

2. **Homo/Heteroscedascity:**

Homoscedasticity is a fancy term for the homogeneity of variance. In the case of the OLS estimation, homoscedasticity in the errors is important to ensure that the least squares estimator is in fact the best linear unbiased estimator. If residuals are heteroscedastic, then a weighted least squares is more prudent in estimation. Fortunately, homoscedasticity is not required to ensure the estimates of parameters are unbiased, consistent, and asymptotically normal; only to ensure the OLS estimator. The intuition is that in prediction, if the errors are homoscedastic, then the error terms do not depend on the independent variable and is therefore easily anticipated.

3. **Least Squares Regression**:

In order to fit a line to our data that has small residuals, we use a method that seeks to minimize the sum of the squared residuals. This is called the least squares method. The notation for the least squares line is : $\hat{y} = \beta_0 + \beta_1 x$ where $\hat{y}$ is our predicted y, β_0 is our intercept, β_1 is our slope and x is the explanatory variable.

The slope of the regression line can be calculated as : $\beta_1 = \frac{sd_y}{sd_x} R$

The intercept is where the regression line intercepts the y-axis.

There are three conditions for a Least Squares Line:

1. Linearity
    a. The relationship between the explanatory and response variable should be linear
2. Nearly Normal Residuals
    a. This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data
3. Constant Variability

a. The variability of points around the least squares line should be roughly constant

4. **r^2, or the coefficient of determination**:

This measurement essentially predicts the proportion of variance of a predicted variable that can be gathered from the independent observation. Mainly used to test a model of the dependent data to confirm findings. In our case, we have simple linear regression, and so this particular model of r^2 is used.

-> Let the series of data points be as follows: x1, x2, …, xn. Then, our residuals are ei = xi-fi (where fi is the value fm the predicted model). Then, the data can be represented using three sums of squares formulas (total, regression, and residuals) where the coefficient of determination becomes (1 - (sum of residuals)/(total sum)).

We interpret the results as an explanation of the goodness of fit of our model, or how well the regression line actually fits the given points for the data.

Works Cited

1) https://en.wikipedia.org/wiki/Coefficient_of_determination - To understand how to interpret the coefficient of determination is used.
2) https://www.youtube.com/watch?v=flnEw5LTvxM&t=902s- Linear Regression in R
3) https://math189.edublogs.org/files/2016/02/chp6-uz9m4a.pdf - For remaining theory